JPRS: 6445

21 December 1960

FOREIGN DEVELOPMENTS IN MACHINE TRANSLATION AND INFORMATION PROCESSING

- USSR -

Distributed by:

OFFICE OF TECHNICAL SERVICES
U. S. DEPARTMENT OF COMMERCE
WASHINGTON 25, D. C.

U. S. JOINT PUBLICATIONS RESEARCH SERVICE
1636 CONNECTICUT AVENUE, N. W.
WASHINGTON 25, D. C.

F O R E W O R D

This publication was prepared under contract by the UNITED STATES JOINT PUBLICATIONS RESEARCH SERVICE, a federal government organization established to service the translation and research needs of the various government departments.

JPRS: 6445

CSO: 3901-D/36

FOREIGN DEVELOPMENTS IN MACHINE TRANSLATION AND INFORMATION PROCESSING

- USSR -

FOREWORD

This translation series presents information from foreign-language literature on developments in the following fields of language data processing. Machine translation studies; questions on structural linguistics, phonological theory, investigation of morphological models, development of syntactic structures and transform analysis; theory of language communication; logical and linguistic problems of constructing information machines and information languages; logical semantics; mathematical and applied linguistics; automatic programming; organization, storage and retrieval of information; automatic indexing and abstracting; character and pattern recognition; automatic speech input; documentation, etc. The series is published as an aid to U. S. Government research.

Previously issued JPRS reports on this subject include:

JPRS: 68, 241, 319, 355, 379, 387, 487, 621, 646, 662, 705, 729, 863, 893, 925, 991, 992, 1006, 1029, 1130, 1131, 1132, 1133, 3225, 3300, 3356, 3433, 3474, 3502, 3532, 3570, 3597, 3598, 3599, 3613, 3629, 3731, 3758, 3796, 3797, 3948, 6094, 6135, 6152, 6236, 6239, 6240 and 6347.

Table of Contents

I. CONFERENCE ON SPEECH STATISTICS

/Following is a translation of an article by V. A. Uspenskiy in the Russian-language periodical Voprosy yazykoznaniya (Problems of Linguistics), Moscow, No. 1, 1958, pages 170-173./

The conference, called by the Speech Section of the Acoustic Commission of the Academy of Sciences USSR and the Leningrad University, held from 1 through 4 October 1957 in Leningrad, was devoted to the subject of speech statistics. Participants included representatives of the Moscow State University, MGPIIYa, the NII of the Ministry of the Radio Engineering Industry, the Institutes of Physiology and Linguistics of the AS USSR, the Laboratories of Electronic Computing VINITI of the AS USSR, the A. F. Mozhayskiy Air Force Academy, and several other organizations.

The conference was opened with an introduction by L. R. Zinder. Attention at the conference was focused on two fundamental problem areas: 1) Application of statistical research of speech and writing for solving problems raised in those realms of contemporary technology which involve in one way or another the storage, treatment, and transmission of information, and 2) the relation of structural and statistical methods in linguistics.

The report "The Meaning of Statistical Research for Technology" was given by L. A. Varshavskiy, who pointed out the fundamental directions in the study of speech, the development of which are essential for telephony (wire and wireless). The first among such directions is the study (including statistical study) of the physical characteristics of sound and electrical signals which transmit speech. In this connection it is expedient to develop the general theory of signals, as pointed out by N. A. Zhelezov in his report, "Power Characteristics of Correlation Intervals of Electrical Signals, in Particular, Speech Signals."

Another problem area, pointed out by L. A. Varshavskiy, concerns the perception of speech transmitted through a communications channel. Channel quality is measured by its articulation, which in communications engineering is the percent correct perceptions of sound units transmitted through the channel. To determine articulation, special sound combinations are transmitted over the channel. When tables of such sound combinations are compiled, statistical relations must be accounted for. But the greatest interest is in the connections between the articulations of sound units of different orders: separate sounds, syllables, morphemes, words, etc. Here the deciding role is played by the effect of the perception of adjacent sounds on the perception of a given sound. This effect is caused by the probability dependence (correlation) occurring between neighboring sounds.

The report of L. R. Zinder, "Linguistic Probability," was devoted to the study of such correlation between neighboring elements of speech (i.e., sounds following each other, words, etc.). As noted in the report, each element of speech carries specific information (in many cases a great deal) on the element immediately following it. Zero order linguistic probabilities (i.e., absolute probability of the incidence of certain elements), as a rule, do not coincide with linguistic probabilities of the first order (conditional probabilities of certain elements occurring after others). Zero and higher order linguistic probabilities are divided into lexical (the probability of occurrence of certain lexemes), grammatical (the probability of occurrence of certain sound units), and sonic (the probability of occurrence of certain sound units). Linguistic probabilities have a significant influence on perception and, in the final analysis, on understanding of speech (this refers in a larger degree to grammatical and to a lesser degree to grammatical and to a lesser degree to sonic probabilities). L. R. Zinder's report was accompanied by a demonstration of three tables of first order sound probabilities for the Russian language. Tables which were prepared on the basis of a statistical study of a text of approximately 90,000 phonemes were shown. These were individual tables showing sound combinations within words and at the junctures of words, and there was a summary table.

Ye. V. Paduchevaya's report, "Statistical Study of Syllable Structure (In Connection with the Use of Information Theory Methods)," was devoted to the comparative analysis of phoneme combinations within words and at word junctures. In this reporter's opinion, only phoneme combinations within a syllable can be a subject of phonological study; combinations at syllable junctures is more appropriately the consequence of word formation. It is natural to expect that the limitations to the combining capability of phonemes within a syllable will be weaker at syllable junctures. At the same time, as this reporter showed with an example from Spanish, a simple division of phoneme combinations into possible and impossible ones is insufficient; probability concepts lead to much more satisfactory results (for Russian).

The most important problems of those enumerated in the report by L. A. Varshavskiy concern compression for the most effective use of communication channels. One of the means of improving the carrying capacity of a channel is the shortening of the time required to transmit one sound. The question arises as to what limit a sound can be shortened. As M. F. Derkach showed in his reports "Statistical Duration of Voiceless Consonants and Their Comprehension" and "Statistics of Characteristic Sound Portions of Vowels in Russian," the acoustical structure of sounds is non-uniform in time. Therefore, if from a time interval during which a sound is made a smaller interval is extracted, comprehension may

be lost. Participants at the conference had the opportunity of listening to a tape recording of the syllable ta which was artificially derived from the syllable sa by clipping the beginning of the sound s (but clipping the end of the sound s in the syllable as does not destroy comprehension).

A more radical way of improving the transmission capacity of a channel would be the extraction of the minimum information necessary to differentiate units of speech (phonemes, for example) and transmission of only this minimum information. (The leading importance of studies to determine the semantic burden of speech units, primarily phonological studies, is obvious.) It is envisaged to place at the transmitting end of the communication line an analyzer capable of differentiating phonemes and transforming them into discrete signals, and to have at the receiving end a synthesizer capable of transforming the signals received into sounds. Fundamentally more complicated is the construction of a differentiator. One of the possible designs of such a unit would be a device capable of determining the phonetic affiliation of a sound according to its spectrogram.

A new method, differing from that usually applied to formant analysis, was advanced by L. A. Chistovich in her report "Application of Statistical Methods to Determination of Phonetic Affiliation of an Individual Vowel Sound." She declined to venture into spectrograms of invariants, which would permit exact determination of the phonetic affiliation of a sound; her idea is that the question of phonetic affiliation of sound can be answered not categorically, but with a certain probability. Spectrogram envelopes of various sounds spoken by different persons were examined. The envelopes for each sound (different for different persons) are averaged, and the result is a basic curve for the given sound. This basic curve is considered to be the signal, while the real envelope (of a speaker) is considered to be the signal distorted by some noise introduced by the speaker. In order to determine the phonetic affiliation of a real spoken sound, the envelope of its spectrogram is taken and its deviation from the basic curve is determined. Probabilities are calculated of whether a given real envelope is a distortion of one of the basic curves or, which amounts to the same thing, whether a given real sound is the sound a, o, u, etc. Chistovich's method can be used without change to determine the phonetic affiliation of sounds in connected speech, accounting for the effect on their comprehension by the comprehension of neighboring sounds.

In his report, "Use of Statistical Methods in Experimental Phonetic and Psychological Study of Speech," V. A. Artemov described the work being conducted at the Laboratory of Experimental Phonetics of the MGPIIYa.

Highly important in telephony (wire and wireless) is the statistical study of writing, especially calculation of the entropy of the distribution of letter combinations. As described in their report, "Statistics of Three-Letter Combinations in Russian Printed Text," by V. A. Garmash and D. S. Iebedev, the Laboratory of the AS USSR for the solution of scientific problems in wire communications computed the frequency of three-letter combinations in a portion of text from Tolstoy's War and Peace. The length of the text was 30,000 letters (the spaces between words were counted as letters and were designated by hyphens). Frequencies of several combinations were given in the report. The most frequent were combinations of "- i -" (frequency 82 x 10^{-4}) and "- ne" (frequency 71 x 10^{-4}). Calculation of the corresponding entropy showed that for suitable coding of three-letter combinations the quantity of text can be decreased by about 5/3 times.

The role of statistical methods in linguistics is not determined exclusively by the requirements of technical applications. In the report "Relation of Structural and Statistical Methods in Linguistics" I. I. Revzin pointed out that it would be quite erroneous to underestimate the value of probability and statistical concepts for the development of linguistic theory itself (if only, in the speaker's opinion, because the structure of language, as a code, is greatly burdened with language redundancy; this redundancy is imposed by the necessity for speech to be resistant to noise). As Revzin pointed out, the usual linguistic methods must be complemented by statistical ones.

In the report "Sequential and Functional Relations of the Vietnamese Verb," by Yu. K. Lekomtsev, it was shown that in Vietnamese, after segregating certain full verbs, auxiliary verbs can be identified on the basis of purely statistical counts of their combinations with full verbs.

Statistical methods are also successfully applied in achieving new linguistic advances. For example, I. A. Mel'chuk proved in his report "Application of Statistics to the Problem of the Category of Gender in French and Spanish" that in French the grammatical category of gender is formally expressed by the noun ending. The author offered the following statistical criteria for the expression of the category of gender: the category of gender is expressed in the given language if there are rules, not too many (not exceeding, let us say, the number of similar rules in Spanish, where the expression of gender is usually not subject to doubt), which permit recognition of gender by noun endings and encompassing no less than 94% of the nouns of the language (these rules include over 98% of the nouns in Spanish).

As noted at the conference, statistics can be used in textology to determine authorship.

In the report, "Probability Determination of Linguistic Time (In Connection With the Problem of Using Statistical Methods in Comparative-Historical Linguistics)," Vyach. Vs. Ivanov demonstrated the possibility of using statistical methods in internal reconstruction. Analyzing the state of a language at a given moment it is possible to establish the direction of linguistic time. For example, linguistic units frequently occurring in texts have the tendency of occurring in subsequent states of the language if their degree of isolation in the system is slight; however, if these frequent units have a high degree of isolation, then they are characteristic of preceding states. Frequency of occurrence and degree of isolation are determined statistically. Another method of reconstruction and prediction is connected with statistical study of linguistic styles. The kinship of languages is most reliably determined in terms of isomorphism of the systems being reconstructed. With reference to coincidence counts frequently used to establish the degree of kinship between languages, as the reporter pointed out, a rigorous probability analysis is needed (since the probability of random coincidence can be high, but meaningless).

I. I. Revzin pointed out in his report the need to develop the special branch of linguistics -- statistical linguistics. Relations have already been found in statistical linguistics for such values as: word frequency, word rank (sequential number in frequency dictionary), word length, etc., which shed light on the theoretical-informational nature of language as a code. This subject was covered by R. G. Piotrovskiy in his report "Certain Problems in the Statistical Study of Lexical Groups."

I. I. Revzin explained in his report the two-sided character of the relation between structural and statistical methods. Statistics not only helps to better understand the structure of a language, but the units themselves which are counted require an accurate structural determination. Thus, in the author's opinion, a deficiency in glotto-chronology is the lack of a precise determination of cognate words. (V. V. Ivanov indicated in his report that there was not enough motivation in even the selection of the basic glossary.) It is evident that statistical study of syllable structure cannot be done without a precise definition of the term "syllable" (in the first part of her report E. V. Paducheva gave such a definition for a Spanish syllable, which, apparently, makes it possible to divide a word precisely into syllables). The importance of structural categories is revealed in statistical studies, which have direct practical application. In order to establish the optimum rules for machine translation, it is necessary to make a statistical study of the languages of the individual sciences. In the report "The Statistical Vocabulary of Russian Mathematical Texts," I. A. Mel'chuk, T. N. Moloshnaya, A. L. Shumilina, Z. M. Volotskaya and I. N. Shelimova presented the results of a statistical study of the language of mathematical literature. In the study

it was necessary to define precisely such concepts as "syntagm," "syntagm type," "word relations in a sentence," etc.

Without doubt, this Leningrad conference was extremely important and not limited to the problems indicated in its title. Two situations were clearly delineated at the conference.

1. The penetration of mathematical, particularly statistical, methods into linguistics is clearly fruitful. These methods can play a very important, though subordinate, role in the solution of linguistic problems. Complete formalization of a real language according to some mathematical system, apparently, can never be achieved; however, there is a possibility of some approximate formalizations of a real language. The deviations between the real language and the approximation must be evaluated statistically.

2. Linguistic studies are acquiring ever greater practical meaning, not limited, as earlier, to school grammars and orthographic rules. This does not mean that linguistic study is losing its theoretical profile. On the contrary, with the development of technology it happens that the finest theoretical constructions are the most important in application. The status of matters in linguistics in this regard is comparable to that in mathematics, where theoretical offshoots (such as mathematical logic) have recently acquired special importance in application.

Especially rewarding was the variety of specializations represented at the conference, from radio engineering to physiology. The conference demonstrated the necessity and further coordination of the work of the representatives of various sciences in the realm of applied linguistics.

II. ON LINGUISTIC PROBABILITY

/Following is a translation of an article by L. R. Zinder in the Russian-language periodical Voprosy yazykoznaniya (Problems of Linguistics), Moscow, No. 2, 1958, pages 121-125./

/Note: Superscribed numbers refer to notes appended to this article./

To know a language means to have in one's memory its words and its grammatical models as determined by its grammatical structure. Words are retained in the memory as sound images, while literate people retain them in the form of visual images. The latter are correlated with sound images if we are dealing with languages written with letters, or are not correlated with them if dealing with languages written with hieroglyphics. The sound aspect of individual words consists of a certain number of repeating elements in various combinations. These elements are called the sounds of speech or phonemes, and in each language comprise a unique system.

Since individual words can consist of various numbers of phonemes, sometimes a large number of them, and though limited in every language[1] the combinability of phonemes is still quite large, hundreds of thousands of words can be formed in a language from a few dozen phonemes. The same can be said of the written form of a language and its system of letters.

Grammatical models, including not only models of sentences, word combinations, and word changes, but also word formation models, do not exist in language separate from words; they exist in some form of word (i.e., in the variations of their sound patterns), in combination with so-called "auxiliary" words (prepositions, conjunctions, etc.), or in some arrangement of words. Furthermore, grammatical models are abstracted from words in language and exist in the form of rules somewhat independent of words. Correspondingly, they are retained in the memory in an abstracted, "schematic" form.

Perception or understanding[2] of speech consists of perception of sound signals in oral speech or visual signals in the case of written matter and their identification with the sound or visual word images reposing in the memory.[3] The richer the memory of an auditor in such images, the more rapid and easier his comprehension of speech. An individual's active vocabulary, that is, the number of words used by him, is relatively small. The number hardly exceeds 10% of the words in a given language (at least of the highly developed languages). An individual's passive vocabulary, i.e., the number of words the meanings of which are known, is considerably larger than the active vocabulary, but even this fund of words is undoubtedly far from all of the words in the language.[4]

It should be noted that because of word formation patterns in the memory there are more "understandable" words in the memory than words. For example, the meaning on first hearing of the word "kholodil'nik" /refrigerator/ becomes clear from the suffix -il'nik (model: "kipyatil'nik" /kettle/). The approximate meaning of the word "vertolet" /helicopter/ is recognized from the model "samolet" /airplane/, etc. A certain number of unknown words, therefore, can be deciphered by a hearer. However, in speech words occur which remain not understood. The development of an individual's vocabulary necessarily passes through the stage of conversion of unknown words into known words.

All of the discussion above permits us to distinguish between perception and understanding. The former consists of recognizing the sound image of a word, the latter consists of connecting this image with meaning. Perception, as we have seen, precedes understanding.

The possibility of perceiving individual words (but by no means speech as a whole) without understanding them results from the fact that there reposes in the memory of the hearer, abstracted from the words, system of sound elements (phonemes) of the language. The recognition of the sound image of a word occurs through these elements.

The distinction between perception and understanding serves as the basis for the long-used system of syllabic articulation in telephony in which meaningless syllables, the perception of which as proved in practice are not subject to any doubt, are used to test communications lines. Further, despite the fact that understanding is excluded here, the perception of sounds (phonemes) comprising such sound combinations, is substantially different from perception of any other sounds in nature.[5] This can be explained by the fact that the composition of language phonemes, which independently do not have meaning, in the final analysis is determined by the meaningful relations existing in the given language. Therefore we can say that although perception of speech sounds, phonemes, and their combinations is possible without understanding, the two are connected.

Understanding noticeably affects perception, makes it easier and more likely for the hearer. This capability of understanding to give additional information on the material being transmitted (whether a sentence, word, or morpheme), together with the probability limitations in the language, determine what in information theory is called "redundancy."[6] Redundancy, as M. P. Dolukhanov writes, "shows the relative quantity of excess information, which is determined by the structure of the language and is known on the receiving end from statistical data."[7]

It is easy to see that in general redundancy is the same as the "psychological factor" or "guess factor," which was spoken of

earlier in telephony.[2] Seeking data on "pure perception" during acoustic testing an effort was made to exclude this factor, and this was the reason for using meaningless sound combinations. Information theory, consequently, has not discovered the phenomenon itself, but it first defined its essence and second provided a unit for measuring it.

It is perfectly obvious that measurement of the effect of the psychological factor is impossible, because it is a subjective factor. An individual's ability to guess depends upon his education, psychic state, degree of fatigue, etc. However, the possibility of guessing is determined objectively by regularities inherent in a given language.

The quantity of redundant information which facilitates guessing, and thereby understanding and perception of speech, is determined by the degree of probability of redundancy in the given language.

Probability limitations characterizing a language are conveniently called the linguistic probability. This term encompasses a variety of phenomena: the probability in speech of individual elements (words phonemes), the probability in speech of grammatical models; the probability of sound combinations and word combinations; in other words the probability of certain sounds following other sounds, and certain words following certain others. All of this is evident to one degree or another in the so-called statistical structure of the language.

The probability of word combination is subject to two different regularities; consequently we can speak of lexical and grammatical probabilities of word combinations. Lexical probability concerns the combination of words depending on their meaning. This dependence, strictly speaking, is not linguistic; it is determined objectively, by social conditions, and is therefore quite variable both geographically and in time. For example, the word combination "sidet' na kryshe" /sitting on the roof/ under conditions of life in the European part of the USSR is immeasurably less probable than the combination "sidet' v sadu" /sitting in the garden/; in central Asia, however, the probability of the first combination increases considerably. In the same manner the word combination "steklyannaya skovorodka" /glass skillet/ was almost improbable twenty years ago, but now it is used quite widely.[8] Thus, lexical probability is a purely statistical regularity independent of the grammatical formation of the language.

Grammatical probability can be of two types. The first type is the probability of a certain part of speech following a certain word; the second is the probability of a certain grammatical form following a given grammatical form. The first type is especially important in languages with rigid word order (English, for example), in which a purely grammatical regularity, a definite grammatical rule, operates in this fashion. It also plays a certain

role in languages having less rigid word order (Russian, for example). Thus, the combination "kholodnaya zima" /cold winter/, in which the noun directly follows its modifier is more probable than the word sequence "kholodnaya nynche zima" /cold winter now/, where the noun and the modifier are separated by a third word, the adverb. The probability in this case is determined not only grammatically but also statistically.

The second type of grammatical probability is especially important in languages having a developed flexional system (Russian, for example) in which rules of agreement and government determine the occurrence of appropriate grammatical forms. In the case of agreement, a purely grammatical rule operates: the word "kholodnaya" /cold/ can only be followed by a singular feminine nominative. In the case of government, statistical rules can also appear if government oscillates, as a rule the words "ne chital" /did not read/ will be followed by the genitive ("ne chital knigi" /did not read a book/), but the accusative ("ne chital knigu") is not impossible.

The quantity of redundant information resulting from grammatical probability is in general very large, especially in cases in which we deal with agreement, and partly also with government. Each of the words in agreement carries information concerning the grammatic form of the other word, that is, information on the grammatical model of the entire combination. It is sufficient to hear the word "deti" /children/ to know that the predicate will be in the 3rd person plural; hearing the attributive "bol'shogo" /large/, we know that the qualified word is a masculine or neuter singular noun. If for example the word "uvidel" /saw/ precedes this attributive, we obtain still more information on the qualified word, which in the given case will almost necessarily be a masculine noun (and denoting an animate object) in the accusative case and functioning as an object. It should be noted that the quantity of redundant information will be all the greater the more universal the pertinent grammatical rule.

The quantity of redundant information resulting from lexical probability is much more varied because it depends entirely on the statistical structure of the language. In the word combination "krasnyy flag" /red flag/, for example, the word "krasnyy" /red/ in some measure contains the information that the word following it is "flag." Furthermore, the quantity of redundant information, as mentioned above, is determined by the probability of occurrence of the combination "krasnyy flag," as compared with the probabilities of combinations "krasnyy tsvet" /red flower/, "krasnyy mak" /red poppy/, etc. Increase in the number of words in the word combination increases the quantity of redundant information. For example, in the word combination "na balkone visel krasnyy flag" /a red flag hung on the balcony/ the probability of occurrence of

the word "flag" after "krasnyy" naturally, is much greater than in the combination "krasnyy flag," because after the words "na balkone visel" it is impossible to have the combinations "krasnyy tsvet," "krasnyy pol" /red floor/, "krasnyy shtab" /red staff/, "krasnyy nos" /red nose/, etc. Corresponding to this, the quantity of redundant information contained in the word "krasnyy" in the first combination is considerably greater than in the second.

The forms of linguistic probability examined are connected with understanding of speech, which assumes that some elements of speech (words or combinations) are correctly heard and understood. Equally with this, perception itself, which, as shown above, breaks down in the final analysis to the perception of individual sound units, phonemes, is facilitated by linguistic probability. In the given case we are speaking of probabilities of compatible sound combinations; the probabilities that a certain phoneme will follow a given one. This probability, which could be called sound probability, is determined in two ways: first, by the phonetic rules of a given language (in Russian, for example, a voiced consonant can be followed only by a vowel or a voiced consonant), second, purely statistically, as with lexical probability.

Since sound probability does not directly affect understanding, its importance as a source of redundancy is not as obvious as that of lexical probability. Nevertheless, sound probability, which exerts a real influence on perception, must also be recognized as a source of redundant information. To illustrate this circumstance we can cite the following fact. As a result of special tests conducted in 1949 it was found that the articulation of the consonants "f," "k," "s" in the sound combinations "fs" and "sk" is greater than the articulation of these same consonants not in consonance (i.e., in combination with other consonants). In one test with a distortion channel the articulation of s out of consonance was 66.6%, in combination with "f" it was 89.8% and in combination with "k" it was 94.7%.[9]

The divergence between articulation of consonants in and out of consonance was explained earlier by the general phonetic position that the acoustic characteristic of an individual consonant and the same consonant in consonance must be different. However, it was not clear why the articulation of a consonant in consonance was always greater than under other conditions. This can be argued, obvicusly, only from the standpoint of information theory; namely, in the sense that a neighboring consonant in Russian contains more redundant information on a given consonant than a neighboring vowel. With regard to consonance combinations "fs" and "sk," they have relatively high probability as compared with other consonantal combinations. This can be used to explain the sharp increase in their articulation observed in the tests mentioned.

In the same manner as in lexical probability, the increase in the number of sound combinations is accompanied by an increase in the quantity of redundant information relative to each of the sounds in the combination. It is perfectly clear, for example, that the occurrence of "r" after "ost" is much more probable than after "s" and even after "st." Consequently "ost" contains much more redundant information relative to "r" than "st," and "st" in turn, more than "s." Thus we can differentiate sound probability of the "first order" when combinations of two sounds are considered, of the "second order" when combinations of three sounds are considered, etc.

A project was carried out in 1956 at the Chair of Phonetics at the A. A. Zhdanov Leningrad University to determine the sound probability of the "first order" in Russian. Raw material, in the light of the ideas discussed above, was gathered by taking articles from current newspapers as well as works of contemporary Soviet authors (Gorbatov, Gaidar, Tendryakov, Paustovskii, Panova, etc.). The total material consisted of 88,538 phonemes. After phonetic transcription of all texts, a count was made of each combination possible in Russian, and then the probability of each combination was calculated.

The sound probability within words is determined by their frequency of use and their morphological structure; at word junctures it is determined by the combining ability of words and by syntactic rules, i.e., by lexical and grammatical probability. As a result, two tables were prepared.

In addition to the forms of sound probability indicated, a determination was made of the sequential probability of individual groups of sounds following each other within words: vowels and consonants, stop consonants and sonants, plosives and fricatives, nasals and liquids. It was found that the probability of occurrence of a vowel after a consonant is 0.7449, whereas the probability of a consonant following a consonant is 0.2551; the probability of a vowel following a vowel is 0.0017, and a consonant following a vowel is 0.9983.

The accompanying tables show the sequential probability of different groups of consonants. The unit for the first table is "consonant + consonant," and for the second a combination of "consonant + any sound (vowel or consonant)." (The letters in the tables are identified as follows: C, consonant; M, voiced consonant; S, sonant; E, plosive; F, fricative and affricative; N, nasal; L, liquid (including j).)

TABLE 1

Preceding	Following C	M	S	E	F	N	L
C	1.0000	0.5695	0.4305	0.3385	0.2310	0.1800	0.2505
M	0.7850	0.4278	0.3572	0.2732	0.1546	0.1161	0.2411
S	0.2150	0.1417	0.0733	0.0653	0.0764	0.0639	0.0094
E	0.3303	0.0981	0.2322	0.0350	0.0631	0.0420	0.1902
F	0.4547	0.3297	0.1250	0.2382	0.0915	0.0741	0.0509
N	0.0878	0.0471	0.0407	0.0230	0.0241	0.0358	0.0049
L	0.1272	0.0946	0.0326	0.0423	0.0523	0.0281	0.0045

TABLE 2

Preceding	Following M	S	E	F	N	L
C	0.2551	0.1453	0.1098	0.0864	0.0589	0.0459	0.0639
M	0.2003	0.1092	0.0911	0.0697	0.0395	0.0296	0.0615
S	0.0548	0.0361	0.0187	0.0167	0.0194	0.0163	0.0024
E	0.0842	0.0250	0.0592	0.0089	0.0161	0.0107	0.0485
F	0.1161	0.0842	0.0319	0.0608	0.0234	0.0189	0.0130
N	0.0224	0.0120	0.0104	0.0059	0.0061	0.0091	0.0013
L	0.0324	0.0241	0.0083	0.0108	0.0133	0.0072	0.0011

NOTES

1. In Russian, for example, voiceless consonants cannot combine with voiced consonants in the same word, hard consonants cannot precede the vowel "i," etc.

2. It will be shown that these two concepts are not identical, although in the narrowest sense they are interconnected.

3. Inasmuch as for the purposes of this paper the written and spoken forms of speech are completely analogous, subsequent discussion will be confined to oral speech.

4. With regard to active vocabulary there are a few published data. They concern the number of words used in the works of such authors as Shakespeare, Pushkin, etc. As for passive vocabulary, there is no reliable method for evaluating it.

5. L. R. Zinder, "Specific Features of the Perception of Speech Sounds," sb. Vospriyatiye zvukovykh signalov v razlichnykh akusticheskikh usloviyakh /Symposium: Perception of Sound Signals Under Various Acoustic Conditions/, Moscow, 1956, page 69.

6. See M. P. Dolukhanov, Vvedeniye v teoriyu peredachi informatsii po elektricheskim kanalam svyazi /Introduction to the Theory of Transmission of Information Over Electrical Communications Channels/, Moscow, 1955; A. A. Kharkevich, Ocherki obshchey teorii svyazi /Essays on the General Theory of Communication/, Moscow, 1955; S. Goldman, Teoriya informatsii /Information Theory/, translated from the English, Moscow, 1957; G. A. Miller, Language and Communication, New York, 1951.

7. M. P. Dolukhanov, op. cit., page 30. The term "redundancy" should be taken with qualifications, because if noise is present it can be impossible to obtain even the minimum information if there is no redundancy.

8. This statement necessitates the conclusion that lexical probability should be determined for a more or less definite interval of time. Word combinations occurring in the works of authors of the 19th century speak nothing for the lexical probabilities in present-day Russian. The same, of course, applies also to the probability of the occurrence of certain words, as well as individual phonemes.

9. L. R. Zinder, "Russian Articulation Tables," Trudy Voyennoy krasnoznamennoy akademii svyazi im. S. M. Budennogo /Proceedings of the Military "Red Banner" Academy of Communications imeni S. M. Budennyy/, Vol. 29-30, Leningrad, 1951, pages 37-38.

III. SYMPOSIUM ON PROBLEMS OF SPEECH DISCRIMINATION

/Following is a translation of an article by L. R. Zincer in the Russian-language periodical Voprosy yazykoznaniya (Problems of Linguistics), Moscow, No. 5, 1958, pages 151-152./

A symposium was held in Moscow from 26 through 30 May 1958 on problems of speech discrimination. The symposium was sponsored by the Acoustics Commission of the Academy of Sciences, USSR. Participants in the symposium were Soviet and Czechoslovak scientists -- physicists, communications engineers, and linguists. Two problems were discussed: 1) the feasibility of using test tables consisting of meaningless sound combinations (syllables), and 2) the types of sound combinations (syllables) included in test tables.

After an introduction by Academician N. N. Andreyev, the following reports dealing with the first problem were heard: from Czechoslovakia -- I. Slavik (physicist) and I. Vakhek (linguist), "New Methods of Testing Speech Discrimination"; from the Soviet Union -- L. R. Zinder (linguist), "Linguistic Principles for Composing Articulation Tables," B. I. Frid (communications engineer), "Relations Between Syllabic and Word Discrimination," and Ye. Yu. Gurbanov (communications engineer), "Problems in the Practical Application of Articulation Tables."

In their report I. Slavic and I. Vakhek criticized the generally accepted method for testing communications channels by using meaningless syllables. Proceding from the fact that this methodology does not account for conditions prevailing for actual telephone conversations, and basing themselves on the fact that "the essence of language is the extremely close connection between the form of the sound and meaningful content," the authors suggest repudiation of the traditional methodology. Instead, they recommend a new, complex method. Fundamental to this method is testing with word tables. In the authors' opinion such tables make it possible to evaluate equally the quality of transmission and sound form and meaning. Further, the complex method allows for tests using conversations on given subjects, and for special cases (for determination of transmittability of individual sounds), tests using meaningless syllables.

In his report L. R. Zinder defended the opposite point of view. In his opinion testing with meaningless syllables is completely permissible from the linguistic point of view for the following reasons: 1) the sound portion of language (phonemes) is abstracted from the meaning portion, forming a special system having a certain independence; 2) perception of the sound form precedes understanding; 3) articulation tests must give an objective measure of the quality of a channel, i.e., its capability of transmitting

a spoken sound form; 4) insofar as a channel transmits only a sound form and not meaning, the testing tables must therefore contain sound material in a form most free from redundancy (in the information theory).

B. I. Frid reported the results of experiments which show that articulation (i.e., percent correctly received) of syllables, articulation of words, and articulation of phrases have the same dependence. For example, if syllables are transmitted better through channel A than channel B, then in the same measure both words and phrases will be transmitted better through channel A than through channel B.

Ye. Yu. Gurbanov spoke first on the theoretical advantages of testing with meaningless syllables, which are the most free from redundancies, from the information theory point of view; second on the great accuracy of results achieved by this method, especially when high-quality channels are being tested; and third on the greater economy of using meaningless syllables as compared with other methods.

In the discussions, participated in by I. Merkhaut and V. Klimesh of Czechoslovakia and V. A. Artemov, Yu. S. Bykov, L. A. Varshavskii, I. M. Sitvak, N. B. Pokrovskii, A. A. Reformatskiy, M. A. Sapozhkov, V. N. Fedorovich, and A. G. El'snits of the Soviet Union, it was brought out first that it was necessary to distinguish between discrimination, i.e., the capability of identifying language elements the sound form of which is not connected with meaningful content (sounds and syllables), and intelligibility, etc., the capability of identifying language elements the sound form of which is inseparably joined with the meaningful content (words and phrases, for example).

It was brought out further that discrimination is not a direct measure of intelligibility, but there is a direct dependence between them. This made it possible for all participants of the symposium to reach the following agreement. "Upon ascertaining that for the given language the various forms of discrimination and intelligibility are interdependent, it is not necessary to measure the various forms of discrimination and intelligibility when determining the quality of a channel. It is sufficient to measure one of the indicated forms, preferably the one which is least difficult to measure. The results of such measurement provides data on other forms of discrimination and intelligibility."

The following reports were read on the second problem: I. Vakhek, "Discussion of a Method of Compiling Articulation Tables," Yu. S. Bykov, "Comparison of the Effectiveness of Various Systems of Articulation Measurements." I. Vakhek in his report criticized the VKIAS syllable articulation tables for the absence in them of open syllables, unstressed vowels, and combinations of three consonants, which are characteristic in Russian. Yu. S. Bykov spoke

of the need for compiling syllable tables from articulation of continuous speech.

L. A. Varshavskiy, Ye. Yu. Gurbanov, L. R. Zinder, and A. A. Reformatskiy participated in the discussions.

IV. ON THE DEVELOPMENT OF STRUCTURAL AND MATHEMATICAL METHODS OF LANGUAGE RESEARCH

/Following is a translation of an unsigned article in the Russian-language periodical Vestnik Akademii nauk SSSR (Bulletin of the Academy of Sciences USSR), Moscow, No. 7, 1960, page 98./

Structural and mathematical methods of language research (structural and mathematical linguistics) are the theoretical basis for solving applied linguistic problems in contemporary cybernetics (automatic oral control of production units, automation of information service, machine translation and abstracting of scientific and technical literature, construction of information-logic machines, construction of automatic stenographic recorders, increasing the carrying capacity of communications channels, etc.). The application of these methods is of considerable value also in the development of theoretical linguistics.

However, as noted in the resolution of the Presidium /of the Academy of Sciences USSR/, scientific research in this area has not received the attention that it merits. Insufficient development of structural and mathematical methods by linguistic departments is hindering practical important work on the theory and practice of machine translation, on construction of information languages and information machines, logical semantics, and cybernetic applications of linguistics.

The Presidium adopted a number of measures to eliminate this situation. It was decided to reorganize the applied linguistics sector of the Institute of Linguistics and in the structure of the Leningrad division of the Institute to create a group for the mathematical study of language and a group for the structural-typological study of languages. A sector of structural linguistics is being organized in the Russian Language Institute and a sector of structural typology of Slavic languages in the Institute of Slavic Studies. A group for structural typology of Eastern languages is being created in the Institute of Far Eastern Studies, and in the Institute of Chinese Studies, a group for structural-typological study of the Chinese language, and in the Institute of Ethnography, a group for structural signalization, calligraphy, and decipherment. The special attention of the reorganized and newly created sectors and groups is directed to the statistical study of the Russian language and other languages of the peoples of the USSR.

Coordination of the work in the area of structural and mathematical methods of linguistic research in the institutes of the Academy of Sciences USSR is made the responsibility of the Scientific Soviet on Cybernetics.

6030

END